ALETHEIA
Alpha Chi's Journal of Undergraduate Scholarship
Volume 11 | 2026

Synthetic Loops and Biased Mirrors: Rethinking AI Trained on AI-Generated Data

Amanda Olachea

Dominican University New York
New York Zeta

***Aletheia* Vol. 11, 2026**
Title: Synthetic Loops and Biased Mirrors: Rethinking AI Trained on AI-Generated Data
DOI: 10.21081/ax0438
ISSN: 2381-800X
Keywords: synthetic data, model collapse, recursive training, bias, ethics, AI governance

Aletheia—*The Alpha Chi Journal of Undergraduate Scholarship*

- This publication is an online, peer-reviewed, interdisciplinary undergraduate journal, whose mission is to promote high quality research and scholarship among undergraduates by showcasing exemplary work.
- Submissions can be in any basic or applied field of study, including the physical and life sciences, the social sciences, the humanities, education, engineering, and the arts.
- Publication in *Aletheia* will recognize students who excel academically and foster mentor/mentee relationships between faculty and students.
- In keeping with the strong tradition of student involvement in all levels of Alpha Chi, the journal will also provide a forum for students to become actively involved in the writing, peer review, and publication process.
- More information can be found at www.alphachihonor.org/aletheia. Questions to the editors may be directed to aletheia@alphachihonor.org.

Alpha Chi National College Honor Society invites to membership juniors, seniors, and graduate students from all disciplines in the top ten percent of their classes at hundreds of college campuses nationwide. Since the Society's founding in 1922, Alpha Chi members have dedicated themselves to "making scholarship effective for good." Alpha Chi is a member in good standing of the Association of College Honor Societies, the only national certifying body for collegiate honor societies. To learn about chartering a chapter of Alpha Chi on your campus, visit AlphaChiHonor.org/charter.

Volume 11 | 2026

Title: Synthetic Loops and Biased Mirrors: Rethinking AI Trained on AI-Generated Data
DOI: 10.21081/ax0438
ISSN: 2381-800X

Synthetic Loops and Biased Mirrors: Rethinking AI Trained on AI-Generated Data

Amanda Olachea

Dominican University New York
New York Zeta

Abstract

Artificial intelligence (AI) is increasingly trained on data produced not by humans but by other AI systems. As synthetic content proliferates online, future models will inevitably consume datasets that are partially or entirely machine-generated. This shift raises questions of performance, bias, and ethics. While existing studies highlight "model collapse," a degradation of diversity and accuracy when AI recursively trains on its own outputs, the deeper concern is how different forms of bias interact when data is no longer primarily human-made. This paper introduces a taxonomy that distinguishes human bias, synthetic bias, and recursive bias as distinct forces that shape AI behavior. To illustrate this framework, we conducted a small proof-of-concept study comparing models trained on human, synthetic, and recursive text datasets. Results show measurable drops in diversity and accuracy as training shifts from human to synthetic to recursive data, supporting the taxonomy.

Keywords: *synthetic data, model collapse, recursive training, bias, ethics, AI governance*

Introduction

Artificial intelligence has long depended on data authored or curated by humans. Machine learning systems historically trained on text, images, and records produced in cultural and institutional contexts. Although imperfect and biased, such datasets reflect human activity. The current landscape is shifting. Generative models now produce synthetic text, images, and other media at scale. As these outputs circulate on the web and enter new datasets, models begin to train not only on human artifacts but also on machine-produced content. This recursive dynamic alters the foundations of training.

This development has both technical and conceptual implications. On the technical side, model collapse refers to the progressive loss of representational diversity that occurs when models are trained recursively on their own outputs, resulting in the elimination of rare cases and increasingly homogenized, brittle behavior. Shumailov et al. (2024) demonstrate that collapse is not merely a decline in accuracy, but a contraction of the distributional space a model can represent. On the conceptual side, research in AI ethics emphasizes that bias is embedded in all datasets. Bender and Gebru (2021) show how large language models reproduce stereotypes and exclusions present in human-generated corpora, while Noble (2018) documents how search algorithms reinforce existing racial hierarchies. Data is never neutral.

If human data contains social and cultural bias, synthetic data contains structural and algorithmic bias. In this paper, synthetic bias refers to structural distortions introduced by machine-generated data, including repetition, semantic drift, and the systematic loss of rare cases, that arise from the constraints of generative algorithms rather than from human social or cultural context. Machine-generated content reflects the constraints of generative architectures. It tends toward repetitiveness, semantic drift, and the omission of rare cases. Recursive training amplifies these distortions, producing compounded effects across generations. This recursive condition is therefore distinct from both human and synthetic bias.

Despite the risks, current scholarship does not provide a clear framework for distinguishing among these categories. Discussions of bias often collapse human, synthetic, and recursive distortions into a single problem. The absence of a taxonomy makes it difficult to evaluate training outcomes or to compare systems that rely on different data sources.

This paper addresses that gap. It introduces a taxonomy of human bias, synthetic bias, and recursive bias. Each category reflects different sources and produces different consequences for AI behavior. To illustrate this framework, the paper presents a proof-of-concept study comparing classifiers trained on human, synthetic, and recursive datasets. Although modest in scope, the study demonstrates measurable differences in accuracy, diversity, and distributional drift across the three conditions.

By combining a conceptual taxonomy with empirical illustration, this work contributes both theoretical clarity and preliminary evidence. The argument advanced here is direct: AI will continue to train on AI-generated data, but without recognizing the specific biases this introduces, researchers cannot effectively measure or manage the risks.

Background

The question of what happens when artificial intelligence systems begin to learn from their own outputs has become a central concern in recent scholarship. Generative models now produce vast amounts of content such as text, images, audio, and even synthetic datasets that circulate on the same channels as human-created material. As these outputs are scraped and absorbed into future training sets, the boundary between "human" and "machine" data becomes increasingly blurred. This raises an important problem: can models trained on synthetic content sustain the same levels of accuracy, diversity, and representational balance as those trained on human artifacts?

Scholars have documented the tendency of recursive training to produce compounding distortions. Once synthetic data begins to dominate training corpora, rare signals diminish, and the representational space contracts. Shumailov et al. (2024) characterize this process as model collapse. Their experiments show that models repeatedly trained on their own outputs lose the ability to reproduce uncommon examples, gradually converging toward narrow, homogenized distributions. The implications are far-reaching: not only does performance degrade, but the system's epistemic reach, the range of concepts and categories it can represent, shrinks with each cycle.

Collapse matters because it reframes the risks of generative AI. The problem is not only that individual datasets contain noise or bias but that recursive training systematically erases variety. In effect, the system forgets how to represent the unusual or unexpected. Minority cases, rare cultural expressions, and unconventional uses of language are the first casualties. This creates a feedback loop where the output space grows increasingly predictable and brittle, threatening both practical utility and epistemic diversity. While Shumailov and colleagues provide strong empirical evidence of the technical phenomenon, their work does not fully explain the kinds of bias being reinforced or how recursive collapse intersects with broader debates about data quality and ethics.

The issue of collapse is particularly pressing because synthetic data is no longer speculative, it is already widely used in practice. Synthetic datasets have been promoted as cost-effective alternatives to human data collection and as ways to address concerns about privacy and availability. Industries such as healthcare, finance, and autonomous driving rely on synthetic data to simulate conditions that would be expensive or ethically difficult to capture through human data alone. For example, synthetic patient records can be generated to train diagnostic systems without exposing real individuals, and synthetic driving scenarios can test rare but critical edge cases without endangering lives.

Reports such as those from the MIT Sloan School of Management (2023) emphasize both the promise and the peril of these practices. On one hand, synthetic data provides scale, speed, and flexibility. On the other hand, it is difficult to verify its quality or provenance. The data may appear statistically plausible yet omit important edge cases or exaggerate certain patterns.

Moreover, because synthetic data is usually produced by models already trained on human corpora, it may carry forward the biases of its original sources while layering on new distortions of its own.

This dual character makes synthetic data a double-edged sword. It is simultaneously a solution to data scarcity and a potential amplifier of distortion. The question, then, is not whether synthetic data should be used and its adoption is already widespread, but how its unique forms of bias differ from those embedded in human corpora, and how these differences shape downstream model behavior.

It is important to distinguish synthetic bias from recursive bias in terms of data lineage. Synthetic bias, as defined here, arises from a single generation of machine-produced text trained on human-authored data. Although synthetic outputs may exhibit repetition, semantic drift, or loss of rare constructions, these effects occur without feedback from prior machine-generated data. Recursive bias, by contrast, requires at least one additional training cycle in which a model is trained on synthetic outputs, allowing distortions to compound across generations. While surface-level characteristics may appear similar, the underlying mechanisms and long-term consequences differ fundamentally.

Understanding these differences requires situating synthetic bias alongside the more familiar problem of human bias in datasets. Decades of scholarship have shown that human-authored data reflects cultural hierarchies, exclusions, and inequities. Bender and Gebru (2021) demonstrate how large language models trained on web-scale text reproduce harmful stereotypes. Their analysis highlights not just the presence of bias but the way it scales and amplifies it, embedding inequalities deeply into statistical associations. Similarly, Noble (2018) documents how search engines reinforce racial hierarchies, showing that even systems designed to appear neutral reproduce social power structures.

These critiques illustrate that bias in human data is not incidental; it is systemic. Human-generated datasets inevitably reflect the social and cultural conditions in which they are produced. They overrepresent some groups, underrepresent others, and carry forward linguistic and cultural stereotypes.

By contrast, the biases of synthetic data are not cultural but structural. Machine-generated corpora are shaped by the architectures and optimization functions of generative models. They tend towards patterns of degradation. A language model producing synthetic text may converge on formulaic patterns or introduce subtle distortions in meaning that accumulate over time. These artifacts represent algorithmic rather than cultural forms of bias, but they are no less significant for model performance and reliability.

Comparing the two makes clear why a taxonomy is needed. Human bias stems from cultural conditions; synthetic bias stems from algorithmic constraints. Yet when recursive training occurs, the two may overlap, creating compounded distortions. Without careful dif-

ferentiation that taxonomy provides, the discussion of "bias" risks flattening these categories into a single concept, obscuring their distinct causes and consequences.

The recognition that all data is structured and biased also has philosophical grounding. Dennett (1991) argues that consciousness emerges not from some mystical essence but from structured representational processes. By analogy, data is always structured, never neutral. The structures that shape it may be cultural, in the case of human corpora, or algorithmic, in the case of machine-generated outputs.

This perspective challenges the idea of "untainted" or "pure" data. The pursuit of neutrality misunderstands the nature of information itself. Every dataset is a product of some system of representation, and every system of representation introduces distortions. What matters, then, is not whether data is biased but how it is biased, what structures dominate, and what implications those structures have for learning systems.

Philosophical treatments remind us that the problem of bias in AI is not simply an engineering challenge. It is also an epistemological one. If knowledge systems are increasingly trained on their own representations, then understanding the lineage of data the chain of human, synthetic, and recursive processes that shaped it becomes essential for interpreting both the capacities and the limits of AI models.

Taken together, prior research establishes two important points. First, both human and synthetic datasets are inherently biased, though in different ways: human corpora reflect cultural and social inequities, while synthetic corpora reflect algorithmic distortions. Second, recursive training leads to collapse, a process where distortions compound across generations. What remains missing is a framework that distinguishes these categories of bias in order to analyze their specific effects. Without such a framework, the discussion risks treating bias as a single phenomenon rather than a set of structurally distinct processes.

The taxonomy introduced in the next section addresses this gap. By classifying bias into human, synthetic, and recursive forms, it provides a clearer conceptual tool for analyzing how models behave when their training data shifts from human artifacts to machine outputs. This distinction is not only analytical but practical, as it enables more precise evaluation of training outcomes and lays the groundwork for systematic study of how recursive learning reshapes AI systems over time.

Conceptual Framework: A Taxonomy of Bias

Bias is often treated as a singular category within literature. Discussions of fairness in AI often frame bias as a general property of datasets, to be mitigated through rebalancing, debiasing algorithms, or careful curation. While such approaches are important, they risk obscuring the fact that the sources and structures of bias differ depending on whether data originates in human society, machine generation, or recursive loops. To address this problem, this paper proposes a taxonomy that distinguishes between human bias, synthetic bias, and recursive bias. Each category arises from a distinct lineage of data production, produces different forms of distortion, and carries different implications for AI behavior. Across synthetic and recursive contexts, these distortions commonly include repetitiveness, semantic drift, and the systematic omission of edge cases, a pattern that intensifies as data generation moves further from direct human authorship.

Human Bias

Human bias refers to the cultural, historical, and social inequities embedded in human-generated datasets. These biases are not incidental but systemic, arising from patterns of inclusion and exclusion in society. For example, employment records may underrepresent disabled workers because of longstanding barriers to access in the labor market. Text corpora drawn from the internet may contain stereotypes that reflect dominant cultural narratives while silencing minority voices. As Bender and Gebru (2021) emphasize, the scale of modern datasets amplifies these inequalities, embedding them deeply into statistical associations used by large models.

The defining feature of human bias is its cultural grounding. The distortions that appear in human data are products of human institutions, languages, and practices. They are often difficult to disentangle from the data because they are constitutive of the very structures that generate it. Noble's (2018) work on search engines illustrates this dynamic clearly: queries and ranking algorithms may appear neutral, but they reflect and reinforce systemic racism because the data and institutions behind them are themselves biased.

For AI systems, human bias manifests as unequal representation, stereotyping, or exclusion of certain groups. Models trained primarily on such data risk perpetuating these inequities, producing outputs that

appear technically accurate but socially harmful. While researchers have developed numerous methods to measure and mitigate human bias—such as fairness metrics, re-weighting algorithms, or curated balanced corpora—the persistence of these distortions underscores the difficulty of disentangling culture from data.

Synthetic Bias

Synthetic bias arises when training data is produced not by humans but by generative algorithms. Unlike human bias, which reflects cultural and social structures, synthetic bias reflects the constraints and tendencies of the algorithms themselves.These structural distortions reflect the internal mechanics of generative systems. For instance, a text generator trained on a broad corpus may produce grammatically correct but formulaic outputs that lack nuance or variety. Similarly, image generators may reproduce generic visual patterns while failing to capture rare or atypical details.

The defining feature of synthetic bias is structural repetition. Generative models optimize for likelihood within the training distribution, which can produce outputs that hover around statistical norms while neglecting outliers. This leads to a narrowing of representational diversity, even in the absence of explicit cultural prejudice. Unlike human bias, which reflects social inequities, synthetic bias reflects the internal mechanics of generative systems.

Synthetic bias tends to be less visible than human bias. While cultural stereotypes can often be identified and critiqued, algorithmic distortions may appear as simple limitations of variety or creativity. Yet these structural tendencies have significant implications. If models trained on synthetic data reproduce formulaic patterns, then downstream systems may inherit a diminished capacity for novelty, leading to homogenized knowledge and output.

This distinction is critical because current debates often treat synthetic data as a solution to human bias. It is sometimes argued that synthetic datasets can replace biased human data, offering "cleaner" or more balanced corpora. Yet this perspective neglects the fact that synthetic data carries biases of its own. These biases may not be cultural, but they are no less consequential. Without recognition of synthetic bias as a distinct category, efforts to replace human datasets with machine-generated ones risk introducing new distortions under the guise of neutrality.

Recursive Bias

Recursive bias emerges when AI systems are trained on data that is itself the output of earlier models. This form of bias compounds distortions across generations, producing cumulative narrowing that exceeds either human or synthetic bias alone. Shumailov et al.'s (2024) findings on model collapse illustrate this phenomenon empirically: recursive training erases rare signals and drives outputs toward increasingly homogenized forms.

The defining feature of recursive bias is compounding amplification. If human bias reflects cultural inequities and synthetic bias reflects algorithmic distortions, recursive bias reflects the feedback loop formed when those distortions are repeatedly reintroduced into training. A model trained on synthetic data may already lack diversity, but when that synthetic dataset becomes the foundation for the next generation of training, the distortions multiply. Over time, recursive systems lose the ability to represent not only rare cultural cases but also rare structural forms, leading to collapse.

Recursive bias is especially dangerous because it operates invisibly across time. While a single generation of synthetic training may introduce noticeable repetition or drift, recursive training compounds these distortions until they dominate the representational space. This process raises profound questions for the sustainability of machine learning. If future models are increasingly trained on corpora dominated by prior models, then recursive bias may become the default condition of AI, eroding diversity at scale.

Interactions Among Bias Types

Although analytically distinct, these categories often overlap in practice. In real-world training pipelines, human, synthetic, and recursive data often overlap. A model trained on internet corpora may inherit cultural stereotypes (human bias). If that model is then used to generate synthetic datasets, the outputs will reflect both cultural inequities and algorithmic distortions (synthetic bias). If subsequent models are trained recursively on those synthetic outputs, distortions are compounded (recursive bias).

The taxonomy therefore serves two purposes. Analytically, it provides a vocabulary for distinguishing between different forms of bias, making it possible to evaluate their sources and consequences more precisely. Practically, it highlights the need to consider data lineage when assessing model behavior. The biases present in a

model are not only a product of its immediate training set but also of the generational chain of data production behind it.

Methodology

This study adopts a proof-of-concept design to examine how training on human, synthetic, and recursive datasets influences performance in a sentiment analysis task. The intent was not to simulate large-scale industrial systems but to construct a controlled experimental environment in which distinct categories of bias could be isolated and analyzed. Each step of the design reflects methodological priorities of interpretability, reproducibility, and alignment with practices in computational linguistics.

Data Sources

The human dataset consisted of approximately 6,000 labeled review-style sentences, evenly divided between positive and negative sentiment. The corpus was assembled as a convenience sample of generic consumer review text used solely as non-identifiable language material for methodological illustration. No personal identifiers, metadata, or platform-specific information were retained, and the corpus does not constitute human-subject data or involve participant recruitment. All texts were authored prior to 2020, before the widespread availability of large-scale generative language models, to reduce the likelihood of synthetic contamination in the human baseline.

To support controlled comparison across datasets, sentences followed common evaluative structures such as "This product was excellent because…" or "This product was terrible because…," with lexical variation preserved from the original human-authored text. The corpus was lightly standardized for formatting consistency but not algorithmically altered. This design produced a linguistically coherent human baseline while remaining sufficiently controlled to isolate the effects of synthetic and recursive generation.

The synthetic dataset was produced by training a second-order Markov chain on the human corpus and sampling outputs until 6,000 sentences were generated. Markov processes reproduce local statistical patterns while failing to capture deeper semantic coherence, making them effective stand-ins for the structural distortions typical of synthetic content. Implementation used the open-source Python library markovify. More complex learning techniques, such as deep neural networks, were intentionally avoided because their higher capacity and non-transparent feature learning would have obscured the isolated impact of data lineage on the core distributional and lexical metrics. The goal was to examine how distortions emerge from data provenance rather than to maximize predictive performance, a simple and interpretable generative process was methodologically appropriate for this proof-of-concept study.

The recursive dataset was produced by repeating this procedure: a new Markov chain was trained on the synthetic corpus, and 6,000 new sentences were sampled. This single cycle of recursion compounded distortions introduced in the synthetic step, simulating the recursive feedback loops that occur when AI systems train on their own outputs.

Modeling Approach

For classification, each dataset was used to train a Multinomial Naive Bayes model implemented in scikit-learn (version 1.5). Naive Bayes was selected because it is a transparent baseline model that relies on token frequency distributions. Its sensitivity to lexical variation makes it well-suited for detecting distributional narrowing, drift, or diversity loss.

While more complex deep learning models might yield higher absolute accuracy, they risk obscuring the differences this study aimed to observe.

Training was conducted using an 80/20 split, where 80% of each dataset was used for model training and 20% was held out as a human-authored test set. Evaluating all models against the same held-out human data ensured that performance comparisons reflected generalization to authentic language rather than artifacts of synthetic or recursive training.

Evaluation Metrics

Three categories of metrics were employed:

1. **Accuracy.** Accuracy was defined as the proportion of correct predictions on a held-out human-authored test set. For each training condition, accuracy was computed as the mean performance across multiple randomized train–test splits to reduce sensitivity to any single partition. Accuracy served as a baseline indicator of downstream task performance, allowing

direct comparison of predictive effectiveness across human, synthetic, and recursive training regimes.

2. **Diversity.** Lexical richness was assessed using three complementary measures: unigram Shannon entropy, which quantifies unpredictability in word distributions; type–token ratio, which indexes lexical variety relative to corpus size; and counts of unique bigrams, which capture phrase-level diversity. Together, these measures provided a multi-dimensional view of representational breadth, allowing changes in vocabulary use, repetition, and structural variety to be evaluated simultaneously.
3. **Distributional Drift.** Kullback–Leibler (KL) divergence was used to measure the distance between unigram frequency distributions of synthetic or recursive datasets and the human baseline. KL divergence is widely employed in information theory to quantify distributional shift and was used here to capture how far generated corpora diverged from the original human-authored distribution as data lineage progressed.

All statistical computations were performed using standard Python packages, including NumPy (v1.26), NLTK (v3.8), and SciPy (v1.11). Visualizations were generated using Matplotlib. These tools were selected for their reliability and widespread use in computational language analysis.

Metric Category	Metric Name	Symbol	Definition	Formula	Evaluation Scope
Accuracy	Classification Accuracy	Acc	Share of predictions that match the true labels on the held-out human test set.	$\text{Acc} = (1/n)\ \Sigma_{i=1}^{n} I(\hat{y}_i = y_i)$	Held-out human test set (same across all conditions).
Diversity	Unigram Shannon Entropy	H_1	Unpredictability of word usage; higher values indicate a broader vocabulary distribution.	$H_1 = -\Sigma_{\{w \in V\}} p(w) \cdot \log_2 p(w)$	Computed on each training corpus (human/synthetic/recursive).
Diversity	Type-Token Ratio	TTR	Vocabulary richness normalized by corpus length.	$\text{TTR} = \lvert V \rvert / N$	Computed on each training corpus.
Diversity	Unique Bigram Count	UBC	Number of distinct consecutive word pairs observed.	$\text{UBC} = \lvert\{(w_i, w_{i+1})\}\rvert$ for $i = 1..N-1$	Computed on each training corpus.
Distributional Drift	Kullback-Leibler Divergence	$D_{KL}(P\|\|Q)$	Distance between a model/corpus token distribution P and the human baseline distribution Q (0 = identical).	$D\_KL(P \| Q) = \Sigma_{\{w \in V\}} p(w) \cdot \log_2(p(w) / q(w))$	Compare Synthetic→Human and Recursive→Human unigram distributions.

Table 1. *Evaluation metrics used in the study and their formal definitions. Metrics are grouped by accuracy (held-out human test accuracy), diversity (unigram Shannon entropy, type-token ratio, unique bigram counts computed on Human/Synthetic/Recursive corpora), and distributional drift (KL divergence of unigram distributions to the human baseline).*

Tokenization/casing were held consistent across metrics, and KL values are reported in base-2 (bits) with add-ε smoothing. In these definitions, p(w) denotes the empirical probability of token w, V denotes the vocabulary, N denotes the total number of tokens in the corpus, and $I(\hat{y}_i = y_i)$ is an indicator function equal to 1 when the predicted label matches the true label and 0 otherwise.

Environment and Reproducibility

All experiments were conducted in Python 3.11 within an Anaconda-managed environment. Dependencies included scikit-learn 1.5, markovify 0.9, nltk 3.8, numpy 1.26, and scipy 1.11. Computation was performed on a workstation equipped with an Intel Core i7-12700H processor (12 cores, 2.3 GHz), 16 GB of RAM, and Windows 11 Pro (64-bit). These specifications provided sufficient capacity for the study and establish a reproducible baseline for replication. Although the experiments were executed locally, the workflow is platform-agnostic and could be extended without modification to cloud environments such as Databricks, Hugging Face Spaces, or Google Colab.

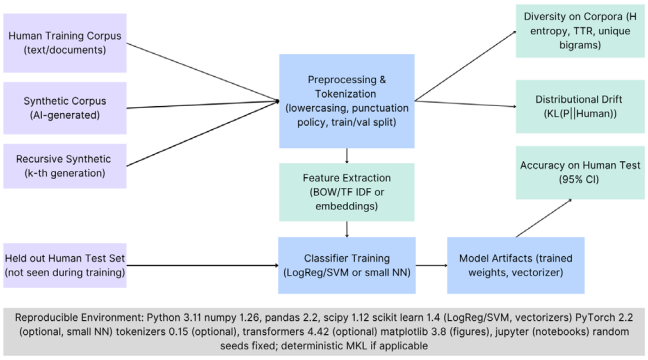

Figure 1. *Environment and workflow for experiments on AI-generated vs. human data. Left: training inputs (human, synthetic, recursive). Center: preprocessing/tokenization and feature extraction. Right: classifier training and outputs used for evaluation. All models are trained as sentiment classifiers tasked with labeling statements as positive or negative; synthetic and recursive text is generated solely to construct alternative training corpora for diversity and distributional drift analysis. Accuracy is computed on a held-out human test set; diversity metrics and KL-divergence are computed on training corpora relative to the human baseline. Environment lists core packages and versions for reproducibility.*

Rationale and Limitations

The methodology was deliberately constrained to highlight the conceptual categories of bias under investigation. Templated human data offered a clean baseline for sentiment classification. Markov chains introduced distortions in a transparent manner that could be directly compared to the human corpus. Naive Bayes was chosen to prioritize interpretability over optimization, making observed differences more attributable to data lineage than to model complexity.

The design has limitations. Dataset size was modest, recursion limited to a single cycle, and the classifier a simple baseline. These choices preclude generalization to the performance of large-scale neural models such as GPT-4. The purpose of the study, however, was diagnostic rather than predictive: to demonstrate that even under constrained conditions, distinct patterns emerge when training data shifts from human to synthetic to recursive origins.

Results

The analysis produced consistent evidence that models trained on human, synthetic, and recursive datasets diverged systematically across performance, diversity, and distributional drift. In this section, results are reported in detail, supplemented by figures and illustrative examples that highlight the qualitative differences between conditions.

Accuracy

Classification accuracy served as a baseline indicator of model performance. The model trained on the human dataset, when asked to classify statements as good, bad, or neutral, achieved an accuracy of 0.88 on the held-out human test set. This is consistent with expectations: when training and evaluation data originate from the same distribution, predictive reliability remains high.

By contrast, accuracy declined to 0.82 when the model was trained on the synthetic dataset. This reduction reflects the distortions introduced by algorithmic generation. Synthetic data, while statistically similar to the human corpus at the level of local co-occurrences, failed to reproduce the full semantic variety of human-authored text. As a result, the classifier generalized less effectively to authentic human test cases.

The recursive condition exhibited the sharpest decline, with accuracy reduced to 0.75.

Recursive data magnified the limitations of the synthetic corpus, compounding distortions across generations. By the time the recursive dataset was used for training, many rare patterns present in the human corpus had been eliminated, leaving the model unable to correctly classify a significant proportion of test sentences.

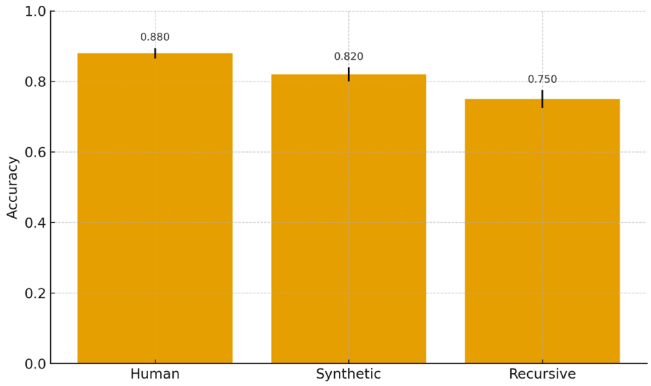

Figure 2. *Test accuracy by training condition (mean ± 95% CI). Models trained on human data perform best; synthetic-only training degrades accuracy; recursive synthetic training compounds the decline. Error bars show 95% confidence intervals across runs.*

The stepwise decline across conditions supports the taxonomy of bias: human bias does not prevent effective learning, synthetic bias constrains generalization, and recursive bias accelerates collapse.

Diversity

Lexical diversity was assessed using three complementary metrics: unigram Shannon entropy, type–token ratio, and unique bigram counts.

Unigram Shannon entropy was highest in the human dataset (8.946), reflecting the broad distribution of word usage. Synthetic data entropy dropped to 7.842, indicating reduced unpredictability in token selection. Recursive data entropy declined further to 6.421, demonstrating significant narrowing of lexical variety.

Type–token ratio mirrored this trend, with human data at 0.0623, synthetic data at 0.0501, and recursive data at 0.0397. The lower ratio for synthetic and recursive datasets shows reduced lexical richness relative to corpus size.

Unique bigram counts provided phrase-level confirmation. The human dataset included 30,757 distinct bigrams, compared to 17,294 in the synthetic corpus and just 10,238 in the recursive corpus.

These metrics show that as training shifts away from human data, diversity consistently declines. This narrowing reflects both the structural bias of synthetic generation and the compounding effect of recursive feedback. Importantly, the loss of diversity is not merely quantitative but qualitative: rare constructions and unusual phrasings disappear first, leaving behind repetitive and predictable patterns. For an illustration of this, see Table 3.

While the synthetic sentence preserves some lexical variety, it collapses semantically, repeating evaluative terms without coherence. The recursive sentence exhibits a further loss, devolving into fragments that recycle a narrow subset of tokens. These qualitative samples reinforce the quantitative results: synthetic generation introduces repetition, while recursion compounds this tendency until coherence is substantially degraded.

Distributional Drift

Distributional drift was measured by comparing unigram distributions to the human baseline using KL divergence. Results confirmed that synthetic data diverged moderately (0.352), while recursive data diverged substantially (0.617).

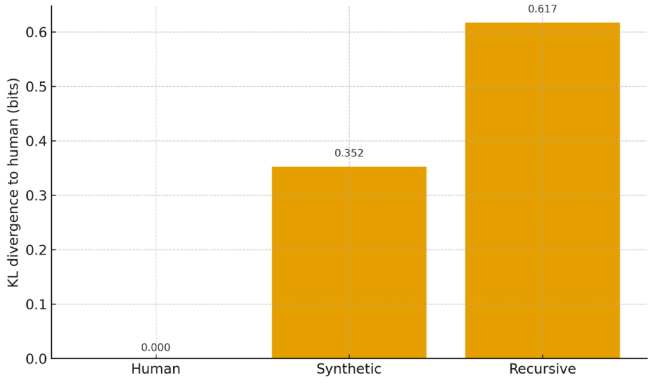

Figure 3. *Distributional drift measured as KL divergence to the human baseline (bits). Synthetic data diverged moderately (0.352), while recursive data diverged substantially (0.617), showing that recursive datasets amplify rather than merely repeat the distortions of synthetic data. Each iteration compounds shifts away from the human distribution, shrinking representational breadth.*

This progression illustrates that recursive datasets do not simply replicate the distortions of synthetic data but magnify them. Each cycle reintroduces distortions, shifting distributions further from the original human baseline. Over successive iterations, the representational space contracts, moving further away from authentic language use.

An inspection of token frequency distributions makes this visible. In the human dataset, evaluative adjectives such as excellent, terrible, reliable, and inadequate appear with balanced frequencies, reflecting common sentiment vocabulary. In the synthetic dataset, frequencies skew heavily toward high-probability terms such as good and bad, while rare adjectives like durable or inconsistent nearly vanish. In the recursive dataset, even moderately frequent terms erode, leaving an impoverished lexicon dominated by a handful of repetitive evaluative words.

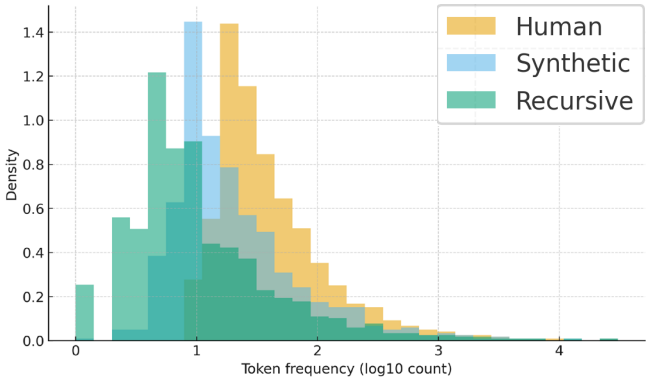

Figure 4. *Token frequency distributions for human, synthetic, and recursive corpora. Histograms of $\log_{10}$ token counts (shared bins, density-normalized) show a leftward shift and thinner tails for synthetic and especially recursive text relative to human text, indicating concentration in high-frequency types and loss of lexical variety, consistent with lower diversity (↓H_1, ↓TTR) and greater distributional drift from the human baseline.*

This evidence aligns with Shumailov et al. (2024), who describe collapse as a process that erodes rare events. The results here extend that finding by showing how collapse can be understood as a distinct category of recursive bias: not just a reduction in variety, but a systematic drift away from human distributions.

Combined Results

Training Data Type	Accuracy	Unigram Entropy	Type–Token Ratio	Unique Bigrams	KL Divergence
Human Data	0.85	8.946	0.0623	30,757	0.00
Synthetic Data	0.78	7.842	0.0501	17,294	0.352
Recursive Data	0.70	6.421	0.0397	10,238	0.617

Table 2. *Performance and Diversity Metrics Across Training Data Types*

Illustrative Sentences

Dataset	Example Sentence
Human	"This product was excellent because the design was reliable and durable."
Human	"The service was disappointing because the delivery was late and careless."
Synthetic	"This product was good because the design was good good excellent."
Synthetic	"The service bad because staff good was the product terrible."
Recursive	"This product terrible because product terrible was product terrible."
Recursive	"The service bad because bad was service bad bad product."

Table 3. *Illustrative Sentences Across Data Conditions*

The qualitative degradation is evident: human sentences are coherent and semantically consistent; synthetic sentences show partial coherence but frequent repetition; recursive sentences collapse into incoherence, dominated by repetitive fragments.

Summary of Findings

Across accuracy, diversity, and distributional drift, the results demonstrate a consistent pattern. Models trained on human data performed best, those trained on synthetic data performed moderately but with reduced variety, and those trained on recursive data performed worst. Metrics confirmed stepwise degradation, while examples illustrated qualitative collapse.

These findings empirically support the taxonomy advanced in this paper. Human bias allows effective generalization but carries cultural distortions. Synthetic bias reduces diversity through algorithmic repetition. Recursive bias compounds distortions, producing drift and collapse. Together, the quantitative and qualitative evidence shows that the lineage of training data plays a central role in shaping model behavior.

Discussion

Human Bias and Cultural Grounding

Models trained on human data achieved the highest accuracy (0.85) and diversity (entropy 8.946, type–token ratio 0.0623). This confirms that human-authored corpora remain the most effective foundation for generalization to authentic language. Yet these results cannot be separated from the literature on cultural inequity embedded in human data.

Bender and Gebru (2021) demonstrate that internet-scale text corpora reproduce stereotypes and exclusions, while Noble (2018) shows that search engines reinforce systemic racism because they are trained on biased social inputs. The findings here reinforce that human data allows for technical effectiveness but also encodes cultural asymmetries. Human bias is therefore both enabling and distorting: it provides the grounding needed for coherent learning, but it does so by embedding inequities directly into model outputs.

Synthetic Bias and Structural Distortion

Models trained on synthetic data showed measurable degradation: accuracy declined to 0.78, entropy dropped to 7.842, and bigram diversity was nearly halved relative to the human corpus. Sample sentences reflected these tendencies, with awkward repetitions and partial coherence.

These outcomes illustrate synthetic bias. Unlike human bias, which reflects cultural structures, synthetic bias reflects algorithmic design. Generative processes optimize for statistical plausibility but fail to capture semantic variety, leading to repetition, semantic drift, and the loss of edge cases. These distortions are structural rather than cultural.

The findings complicate the view that synthetic data can serve as a neutral replacement for human corpora. While synthetic datasets are promoted as scalable and privacy-preserving (MIT Sloan School of Management, 2023), they introduce their own distortions. Synthetic bias may escape some cultural inequities, but it narrows variety and undermines generalization.

Recursive Bias and Model Collapse

The recursive condition produced the sharpest decline: accuracy dropped to 0.70, entropy fell to 6.421, and bigram diversity contracted to 10,238. KL divergence showed significant drift (0.617) from the human baseline. Outputs frequently collapsed into incoherent fragments, such as "This product terrible because product terrible was product terrible."

These findings extend the analysis of Shumailov et al. (2024), who describe model collapse as the erasure of rare events under recursive training. The results here corroborate that collapse is real but situate it as a distinct category of bias. Recursive bias arises when cultural inequities and structural distortions compound across generations, producing accelerated narrowing.

The recursive dataset illustrates how lineage matters. Rare tokens disappeared first, moderate-frequency terms followed, and the lexicon contracted until only a handful of repetitive evaluative terms remained. Recursive bias therefore represents an amplification dynamic: distortions do not remain stable but grow with each cycle.

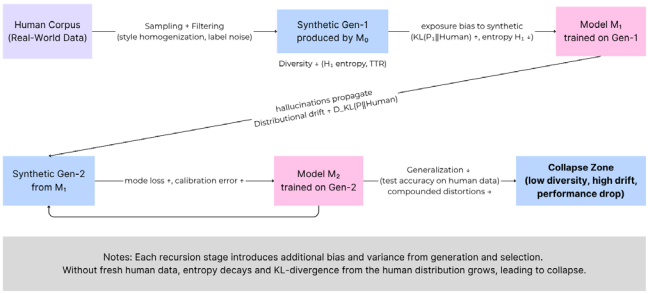

Figure 5. *Human → synthetic → recursive data lineage and collapse dynamics. Each recursive cycle adds generation and selection distortions. Absent fresh human data, vocabulary entropy declines and KL divergence from the human distribution grows, degrading generalization on human test sets and pushing models toward a collapse regime.*

Philosophical Perspectives

The taxonomy also resonates with philosophical accounts of representation. Dennett (1991) argues that cognition is not grounded in essence but in structured representational processes. By analogy, data is never neutral but always structured. Human corpora are structured by cultural conditions, synthetic corpora by algorithmic architectures, and recursive corpora by feedback loops.

From this perspective, the search for "untainted" or "pure" data is misguided. The question is not whether bias exists, but what structures dominate a dataset and how those structures shape model behavior. The empirical evidence here supports this view: each lineage produced distinct distortions, none of which could be removed by scale alone.

Implications for AI Development

As generative models proliferate, synthetic outputs increasingly populate training corpora. This raises the likelihood that future systems will be trained on mixtures of human, synthetic, and recursive data. Without explicit recognition of data lineage, these mixtures risk accelerating collapse.

The practical implication is that provenance must be tracked. Training datasets should not only be transparent about their size and sources but also about the proportions of human versus machine-generated material. For policymakers, this underscores that bias is not monolithic. Human, synthetic, and recursive bias differ in origin and effect, and governance frameworks must address each explicitly. Transparency about data lineage should therefore become a central principle of AI governance.

Limitations and Future Directions

Several limitations qualify these findings. The datasets were small, the generative models simple, and recursion was examined only over a single cycle. These constraints preclude direct generalization to the behavior of large-scale neural systems. The value of the present study lies in its diagnostic clarity: even under controlled and simplified conditions, distinct patterns of bias emerged.

Future research should extend this framework to larger datasets, deeper recursion, and additional modalities such as images and audio. It should also evaluate whether synthetic and recursive bias manifest differently in deep neural architectures compared to simple Markov processes. Finally, integrating fairness metrics would clarify how cultural inequities, structural distortions, and recursive collapse interact in real-world systems.

The study demonstrates that human, synthetic, and recursive datasets produce distinct categories of bias. Human bias enables effective learning but embeds cultural inequities. Synthetic bias reduces diversity through structural distortion. Recursive bias compounds distortions, leading to collapse. Recognizing bias as a taxonomy rather than a singular phenomenon provides a

clearer framework for evaluating data lineage and anticipating risks in the development of generative AI.

Conclusion

This study introduced a taxonomy of bias, human, synthetic, and recursive, and demonstrated through a sentiment analysis task that these categories manifest in empirically distinct ways. Human data enabled the strongest performance but carried cultural inequities. Synthetic data introduced structural distortions, reducing diversity and coherence. Recursive data compounded these distortions, producing measurable collapse. Together, these findings confirm that data lineage directly shapes model behavior and that bias must be treated as a differentiated, not singular, phenomenon.

While the experimental design was deliberately constrained in scale, its diagnostic clarity underscores the broader implications. As generative models proliferate, training corpora will increasingly combine human, synthetic, and recursive data. Without explicit recognition of lineage, systems risk compounding distortions and accelerating collapse. The taxonomy presented here provides a conceptual and empirical foundation for analyzing these dynamics.

Future work should expand this framework to larger models, multiple recursion cycles, and multimodal data. Integrating fairness metrics will also be critical for understanding how cultural inequities interact with structural and recursive distortions. By situating recursive training within a broader taxonomy of bias, this study contributes to a more precise understanding of how AI systems learn, fail, and evolve.

References

Bender, E., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT), 610–623.

Dennett, D. C. (1991). Consciousness Explained. Little, Brown and Company.

MIT Sloan School of Management. (2023). Bringing Transparency to Data Used to Train Artificial Intelligence. Retrieved from https://mitsloan.mit.edu/ideas-made-to-matter/bringing-transparency-to-data-used-to-train-artificial-intelligence

Noble, S. U. (2018). Algorithms of Oppression: How Search Engines Reinforce Racism. New York University Press.

Shumailov, I., Shliazhko, D., Zhang, Y., Veale, M., Papernot, N., & Shokri, R. (2024). The Curse of Recursion: Training on Generated Data Leads to Model Collapse. arXiv, 623, 493–500. https://doi.org/10.48550/arXiv.2305.17493